THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

February 05, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: *60/462,777*
FILING DATE: *April 14, 2003*

> **PRIORITY DOCUMENT**
> SUBMITTED OR TRANSMITTED IN
> COMPLIANCE WITH
> RULE 17.1(a) OR (b)

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

L. Edelen

L. EDELEN
Certifying Officer

Please type a plus sign (+) inside this box → ⊞   04-15-03 60462777-041

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53 (c).

Express Mail Label No.   EL 827 841 037

Date of Deposit: April 14, 2003

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
LALITHA	AGNIHOTRI	FISHKILL, NEW YORK
NEVENKA	DIMITROVA	YORKTOWN HEIGHTS, NEW YORK

☒ Additional inventors are being named on the 1 separately numbered sheets attached hereto

TITLE OF THE INVENTION (280 characters max)

MUSIC VIDEO SUMMARIZATION USING CONTENT ANALYSIS

CORRESPONDENCE ADDRESS

Direct all correspondence to:

☒ Customer Number   24737   →   ***24737***

OR   Type Customer Number here

☒ Firm or Individual Name   PHILIPS ELECTRONICS

Address	580 WHITE PLAINS ROAD
Address	
City	TARRYTOWN
Country	USA

State	NEW YORK	ZIP	10591
Telephone	914-332-0222	Fax	914-332-0615

ENCLOSED APPLICATION PARTS (check all that apply)

☒ Specification Number of Pages | 26 |

☒ Drawing(s) Number of Figures | 15 |

☐ Application Data Sheet. See 37 CFR 1.76

☐ CD(s), Number | |

☐ Other (specify) | |

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT (check one)

☐ Applicant claims small entity status. See 37 CFR 1.27.

☐ A check or money order is enclosed to cover the filing fees

☒ The Commissioner is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 14-1270

☐ Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($)

160.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are: _____.

Respectfully submitted,

SIGNATURE _____   Date APRIL 14, 2003

TYPED or PRINTED NAME   LAURIE E. GATHMAN

REGISTRATION NO.: 37,520
(if appropriate)

TELEPHONE   (914) 333-9605

Docket Number: US 030086

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

S:\ga\US030086Prosheet.doc

PROVISIONAL APPLICATION COVER SHEET
Additional Page

Docket Number	US030086	Type a plus sign (+) inside this box →	+

INVENTOR(S)/APPLICANT(S)

Given Name (first and middle (if any))	Family or Surname	Residence (City and either State or Foreign Country)
JOHN	KENDER	LEONIA, NEW JERSEY

Number 2 of 2

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

S:\ga\US030086Prosheet.doc

United States Patent & Trademark Office
Office of Initial Patent Examination

Application papers not suitable for publication

SN _60462777_   Mail Date _04/14/03_

- ☐ Non-English Specification
- ☑ Specification contains drawing(s) on page(s)_____ or table(s)_____
- ☐ Landscape orientation of text ☐ Specification ☐ Claims ☐ Abstract
- ☐ Handwritten ☐ Specification ☐ Claims ☐ Abstract
- ☐ More than one column ☐ Specification ☐ Claims ☐ Abstract
- ☐ Improper line spacing ☐ Specification ☐ Claims ☐ Abstract
- ☐ Claims not on separate page(s)
- ☑ Abstract not on separate page(s)
- ☐ Improper paper size – Must be either A4 (21 cm x 29.7 cm) or 8-1/2"x 11"
  - ☐ Specification page(s)_____
  - ☐ Drawing page(s)_____
  - ☐ Abstract
  - ☐ Claim(s)
- ☐ Improper margins
  - ☐ Specification page(s)_____
  - ☐ Drawing page(s)_____
  - ☐ Abstract
  - ☐ Claim(s)
- ☐ Not reproducible   Section
  - Reason
  - ☐ Paper too thin
  - ☐ Glossy pages
  - ☐ Non-white background
  - ☐ Specification page(s)_____
  - ☐ Drawing page(s)_____
  - ☐ Abstract
  - ☐ Claim(s)
- ☐ Drawing objection(s)
  - ☐ Missing lead lines, drawing(s)_____
  - ☐ Line quality is too light, drawing(s)_____
  - ☐ More than 1 drawing and not numbered correctly
  - ☐ Non-English text, drawing(s)_____
  - ☐ Excessive text, drawing(s)_____
  - ☐ Photographs capable of illustration, drawing(s)_____

60146277 .04.1403
704117
R7B

INVENTION DISCLOSURE

PHILIPS
GATH
03 APR 10

| RETURN COPY OF THE   PS: 0202 (Please assign to Laurie Gathman   ID # and date in: _____ |
| ID ABSTRACT TO: _BETSY McILVAINE_____   FOR IP&S USE ONLY |

FOR ADMIN USE ONLY

| ECCN: NLR EAR99 | ECO Initials: BM | ECCN Date: 10 Apr 03 | LR/TO: SSY |

INSTRUCTIONS:
1. Description should be supplemented by attaching copies of relevant documents, such as published articles or patents, product brochures, engineering notebook pages and drawings.
2. All inventors must sign the Disclosure of Invention.
3. Please submit completed form to Export Control Officer (ECO), who will assign an ECCN and forward this disclosure to IP&S. Please print single-sided, and do not send duplicates or send directly to IP&S.

Recommendation: Management / patent coordinator as to urgency, commercial importance, and competitors' activities: ☐ LOW   ☐ MED   ☐ HIGH

Title of the invention:

Music Video Summarization using Content Analysis

Name(s) of the inventors	Division Location /ISC Manager's Name	Email/Tel./Job Title	Home Address
Lalitha Agnihotri	HIT/SSA	Lalitha.Agnihotri@philips.com (914) 945-6476 Title: MRS	5 Loudon Dr. Apt. 6 Fishkill, NY 12524 Citizenship: India
Nevenka Dimitrova	HIT/SSA	Nevenka.Dimitrova@philips.com (914) 945-6059 Title: PMRS	3148 Gomer st Yorktown Heights, NY 10598 Citizenship: Macedonia
John Kender	Columbia University	jrk@cs.columbia.edu (212) 939 7115 Title:Professor	169 Ames Avenue Leonia, NJ 07605 Citizenship: US
		@ ( ) Title:	Citizenship:

PRIMARY CONTACT

Who should IP&S contact for further technical information about the invention and its planned use or public disclosure?
Inventor Name: Lalitha Agnihotri

Summary of the invention, where KEY WORDS are underlined which might be useful in searching for relevant patents or publications:

We provide a novel method for **automatic indexing and summarization of complete music video programs and summarizing songs** using multimedia analysis. The steps that need to be performed are **segmenting individual songs** in a music video. We use videotext detection and recognition of overlaid text in the videos along with audio classification, cut-rate variation, superhistogram analysis to achieve this. For summarization of songs, we utilize the chorus of the song. The chorus of the songs is recognized by performing analysis on the transcript of the songs obtained from the closed

INVENTION DISCLOSURE (continued)

captions. The songs are then **identified** using chorus of the song and a database of lyrics. Information regarding the title, artist, awards won, gossip etc. is retrieved from the web about the song and can be displayed in the summary.

Detailed description of the invention on annexes; please describe preferred embodiments and their advantages over prior solutions in detail; please include drawings. (See last 2 pages)

STAGE AND IMPORTANCE OF THE INVENTION:

a.   Stage of the invention?   ☐ Idea   ☐ Pre-development   ☐ (trial) manufacture
                                  ☒ Research   ☐ Development

b.   Was the invention made under a government contract? ☐ No ☐ Yes, contract number: _____

c.   Date of Conception of Invention: _____July 30 2001_____

d.   Date of First sketch, drawing (provide copy if available): _____July 30 2001_____

e.   Date of first written description (provide copy):   _____

f.   Date of completion of first model or full sized device: _____

g.   Date of first successful test: _____

h.   In what products, processes or systems could the invention be used? Web browser, DVD+RW combi_____

i.   For which other business units of Philips could the invention have relevance? _____

j.   For which competitors of Philips could the invention have relevance? Why? AOL Time Warner, TiVo, SonicBlue, ZapMedia

DISTRIBUTION OF INFORMATION CONCERNING THE INVENTION

When, how, where and to whom will information concerning the invention be distributed outside Philips? .
Please consider publications, hearings, exhibitions, offers, contacts with potential customers or suppliers, issuing of samples, trad shows, test sites, public demonstrations, public displays and first offers for sale or commercial use, experimental use outside Philip such as research partnerships, beta tests, regulatory requirements.

Has a description of the invention been, or will be, published or submitted for publication?   ☐ No ☒ Yes
If Yes, provide dates and names of publications: to be submitted to ACM Multimedia 2003, November 2-8, Berkeley CA

N.B. Even after sending this Invention Disclosure to Philips Intellectual Property & Standards, any such acts will impa patentability of the invention. Please contact Philips Intellectual Property & Standards before information concerning th invention leaves Philips.

SUPPLEMENTAL INFORMATION CONCERNING THE INVENTION

a.   Is the invention the result of cooperation with persons outside Philips?   ☐ No ☒ Yes
     If so, with whom? *Prof. John Kender*

b.   Is there, or will there be, an internal report on the invention?   ☒ No ☐ Yes
     If so, please state the number. _____

c.   Are there, or will there be, other invention disclosures relating to this invention?   ☐ No ☒ Yes
     If so, please state Ref. no. US020206, "System and Method for indexing and Summarizing Music Videos," June, 13, 2002.

d.   Are there other persons who could give information on the invention?   ☒ No ☐ Yes
     If so, who? _____

INVENTION DISCLOSURE (continued)

In your detailed description, please indicate:

PRESENT STATE OF THE ART

Briefly describe the closest already-known technology that relates to the invention. This would include, for example, already existing products, methods or compositions which are known to you personally or through descriptions in publications.

Methods to extract closed caption from the stream exist. Also there are disclosed methods to detect text on the screen and to recognize them (A23, 839, SN 09/441,949). Audio methods for chorus detection using auto-correlation exist in the literature.

PROBLEM SOLVED BY THE INVENTION

Briefly describe the problem for which the invention provides a solution. Is this problem new?

We attempt to automatically extract a comprehensive summary of music video programs using multimedia analysis in order to enable users to browse and search music videos. People will have the ability to email summaries to friends in order to recommend new music. It will enable people to create playlists of songs that they like and to find the songs that have been recommended to them. It will enable people to discover new artists and songs.

ADVANCEMENT IN STATE OF THE ART

Briefly describe the unique advancement achieved by the invention. This may be done, for example, by describing a problem with the prior art that is solved or specific objects that are achieved by the invention.

We propose summarization of music videos using multimedia analysis and web infromation extraction.

WHAT IS THE BEST WAY YOU KNOW OF TO IMPLEMENT THE INVENTION?

Briefly describe the invention and how it achieves the advancement described above. Please include at least one embodiment of the invention, with drawings, graphs, test data etc.

(Please Note: If we decide to file an application on this invention, the attorney writing the application will need this information from you in as much detail as possible in order to complete the application.)

See attached.

SIGNATURES

Disclosures must be signed by all of the inventors.

INVENTOR #1: _____ Date: _____

INVENTOR #2: _____ Date: _____

INVENTOR #3: _____ Date: _____

INVENTOR #4: _____ Date: _____

Music Videos Miner

Lalitha Agnihotri
Dept. of Computer Science
Columbia University
New York, NY 10027
+1 212 939 7107

lalitha@cs.columbia.edu

Nevenka Dimitrova
Philips Research
345 Scarborough Road
Briarcliff Manor
+1 914 945 6059

nevenka.dimitrova@philips.com

John Kender
Dept. of Computer Science
Columbia University
New York, NY 10027
+1 212 939 7115

jrk@cs.columbia.edu

ABSTRACT

Music videos currently can be watched 24 hours a day via broadcast video on VH1, MTV1 and MTV2 and a variety of local channels. However, it is difficult to track all these channels and their offerings. People find it difficult if not impossible to find specific videos of interest. It is important to provide tools for browsing, searching and accessing music videos quickly. We present in this paper the digest of user-needs analysis we performed to find out what is important in music summaries. We demonstrate and evaluate a system that summarizes music videos and provides an interactive interface for browsing them. Starting with full music video programs, we segment out the individual song videos by finding their boundaries, which are distinguished by changes in color palette, in closed captions, and in frequency of shot transitions. The video summaries consist of list of song summaries. Each song summary consists of automatically selected high level information such as title, artist, duration, text of the chorus, as well as important audio and visual segments from the input video including the chorus as the most easily recognizable part of the song. Chorus locations are found noting patterns (autocorrelations) of repeated words and phrases in the lyrics. We present the results from a user survey to evaluate the 1) value of the summary 2) content of the summary 3) context of the summary where and how the summary is viewed. We report that this summarization method yields high measurable user satisfaction. Based on the evaluation we designed and implemented a Web based application, called Music Video Miner, which allows people to retrieve music videos by artist, song, and genre.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]. Image/video retrieval

General Terms

Algorithms, Human Factors.

Keywords

Music video summarization, multimedia content analysis, user needs analysis, chorus detection, music databases.

1. INTRODUCTION

Music videos currently can be accessed via broadcast video on VH1, MTV1 and MTV2 24 hours a day. However, a viewer must abide by the broadcaster's terms in order to watch desired songs. A video recorder lets you record whole music shows, however there is no way to specify: I want to watch the song "You are the one" by Shania Twain even on advanced digital recorder products. Content analysis methods have been introduced in the literature that aim at providing high level access to specific parts of the program (e.g. highlights) [Scout]. Video summarization methods have been developed for news, sports, movies, sitcoms. While music content analysis is an active area of research, music videos analysis and summarization has been neglected amongst the existing work. We are presenting a system for browsing and searching music videos based on song summaries. Access to the individual song summaries is via a Web-based interface. Potential applications are numerous for music lovers, special interest groups, music producers, up and coming artists, as well as for copyright infringement detectors.

Video summarization has been a very active area of research [4][8]. However, music videos as a genre has not been investigated before. Reported content based retrieval systems include Query-By-Image-and Video-Content (QBIC) [17], VisualGrep [12], DVL of AT&T, InforMedia [13], VideoQ [7], MoCA [19], Vibe [6], and CONIVAS [1]. The InforMedia project is a digital video library system containing methods to create a short synopsis of each video primarily based on speech recognition, natural language understanding, and caption text. The MoCA project is designed to provide content-based access to a movie database. Besides segmenting movies into salient shots and generating an abstract of the movie, the system detects and recognizes title credits and performs audio analysis. Sundaram et al. [20]proposed a maximization utilization framework for creating audio-visual skims. Agnihotri et al. [2]introduced surface summarization of TV programs using transcript information. Aner et al. introduced mosaic-based scene representation that allows fast clustering of scenes into physical settings, as well as further comparison of physical settings across videos [4]. This enables

detection of plots of different episodes in situation comedies and serves as a basis for indexing whole video sequences. Ma et al. proposed an attention model that includes visuals, audio, and text modalities for summarization of videos [16].

On the other hand, music analysis and retrieval has only focused on the audio aspect [5]. Logan and Chu developed algorithms for finding key phrases in selections of popular music for audio thumbnailing [15]. Their method focused on the use of Hidden Markov Models and clustering techniques on mel-frequency cepstral coefficients (MFCCs), a set of spectral features that have been used with great success for applications in speech processing. Foote introduced audio "gisting," as an application of his measure of audio novelty [10]. This audio novelty score is based on the similarity matrix, which compares frames of audio based on features extracted from the audio. Foote leaves details such as the similarity metric and feature class as design decisions. Many of the previous methods detected as a chorus a repeated section of a given length and had difficulty identifying both ends of a chorus section and dealing with modulations i.e. music key changes. Recently the RefraiD method attempted to detect the chorus sections and estimate both ends of each section [11]. Peeters et al. derive dynamic features representing the time evolution of the energy content in various frequency bands [18]. Their approach is to consider the audio signal as a succession of "states" at various scales corresponding to the structure at various scales of a piece of music using unsupervised learning methods.

The paper is organized as follows. In section 2 we present the user needs analysis for finding a) what is the utility of music summaries and b) what is the important information in the summaries. In section 3 we present a system for automatic music video summarization. In section 4 we present the method of music segmentation and summarization. In section 5 we describe the experimental results and the user tests performed for evaluating the music video summarization method. Music Video Miner, a Web based application for music video browsing is presented in section 6. We conclude the paper in section 7.

2. USER NEEDS ANALYSIS

In order to ascertain the situational utility of music videos summaries, we decided to perform a user-needs analysis. Our test group consisted of eighteen people with ages range from 16 to 53 years. There were eleven women and seven men in our test group. The sessions were conducted one on one and were split into two parts. In the first part, a series of questions were asked which were modified based on the interest the person showed in a subject. After getting the users input on summarization, they were shown a series of summaries and asked to pick the one they liked most. People, could also offer alternate preferable representations if they did not like any of the summaries.

2.1 Q&A Session

The Q&A session was introduced by asking the group if they listen to music and if they watch music video channels (MVC). A series of questions were then asked. The participants were very forthcoming about their ideas and use of such a system. Everybody said that they could use such a system in their life that records and summarizes music videos. In the following section we present the different questions and a summary of the provided answers.

2.1.1 What do you like/do not like about MVCs?

Watching MVC's was almost universal. Viewership was, however, moderated by excessive talk, non-music related shows, or ill-matched choices on the MVC. Everybody enjoyed listening to music and watching the accompanying videos.

2.1.2 Would you like a tool that would allow access to a music videos library instead of watching MVC?

The participants were very excited about this idea and felt that it would be great to be able to watch music videos when they could.

2.1.3 Would you like to see a summary of the videos before viewing it?

All but one answered positively and enthusiastically to this question. Everybody felt that a little bit about the song would be good to see before deciding to watch the entire song. However, one participant felt that as long as the system retrieves all videos by a particular artist that has been requested, then he could just play it.

2.1.4 What should this music video summary contain?

Most answers included the title of the song along with the artist and the album the song came from. The year of the song should also be present in the summary. Seventeen out of eighteen preferred to hear the chorus of the song for the audio summary. Only one person said that the song piece does not matter. One participant said that beginning of the song also gives some idea about the song and would be nice to hear. The summary should include parts with singing rather than just music. The summary could depend on the genre. A shot of the performing artist or the band in the summary would be nice addition. The group felt that the video shown has to be somehow unique and particular to that song. However, for certain genre of songs, the group felt that there could be no distinctive part. Maybe the best person to select the video segment would be the director or marketing representative.

For most of the participants the length of the song is not an issue and hence does not need to be in the summary. Two people said that lyrics of the chorus are not important and should not be included in the summary.

2.1.5 Would you like to get additional information about the song, artist, etc?

There was a wide variation of opinions on this question. Some felt very strongly that they would never want any additional information about the song. Others felt that additional information should be presented as a hyperlink that the users could explore in case they decided that they like the song. Still other felt that additional information would be nice and should be included.
Links could be given for people to find songs in the same album or other songs by the artist/group. A link to discography of the artist was also mentioned. Information about where to buy the song was requested. Picture of CD cover alongside so that they can buy it when they go to a music store was mentioned several times. Statistics regarding how many copies of CD sold, the standing in Billboard charts, awards won would be interesting to see too. The director, location, the actors in the video were mentioned by a few participants who are into music videos.

2.1.6 How would you search such a music videos library?

Search by name of the artist and genre of the song came up most. Searching or retrieval by the most current songs was also a popular choice. Participants want to keep up to date with music and wanted the ability to sort by the most popular songs. They also want the ability to search based on themes, such as rainy videos, relaxing music etc. Ability to search by lyrics of a song they have heard or to search the songs that do not contain the title in the song is important. One person wanted to search for the video that shows a deer driving a car!

2.1.7 Where do you see a good use of such a system?

One recurrent theme for use of such a system was to create party lists. For people with big screen televisions, it is good to get a quick preview of audio and video in order to set up play lists for different parties (Christmas party play list, New Year party play list etc). Another theme was getting the top N songs. People said that they would like to get top 10 popular music videos at the end of the day. One person felt that it would be a good way of discovering new songs and for exploring new genres. People would like to use a system or finding the video that somebody recommended to them. They felt people could use the music videos retrieved by "dance" theme to learn new moves. Another market could be the karaoke kind of application. Here the lyrics have to be matched up to the closed captions.

Ability to get a music channel much like the satellite radio channels that is currently available so that only music videos are shown and there are no commercials and hosts was desired. Another scenario would be for a family where children could select their music while the parents search for their own music preference.

2.1.8 Other comments?

Participants felt that summaries should be different depending on whether they knew the song or not. If they already knew the song, then the title and artist are enough for them to venture into the song and the piece of the audio in the summary does not matter much. However, if they do not know the song, then more information is needed. Participants did not want excessive text, as it requires too much effort. A few people said that music videos are not very important. It is the song itself that is of more value to them. People did not want to pay for such a system.

2.2 User Interface Selection

Once the users answered the questions, they were shown different types of "results" that they might get when they search for a song in our system. The screen shots that were shown to the users are presented in figure 1-5. Once the users picked one out of the five versions, they were shown two more versions of the style that they liked. In one, the image was linked to audio and in the second the image was linked to audio and the video of the chorus of the song. The users then had to choose between the still image vs. audio vs. audio+video version that they would prefer to view. All but one participant had heard the song or recognized it after listening to the audio. Out of the five different presentations, all of them were equally selected. Most people did not have a strong preference about one presentation or other. However, some people preferred getting the shorter version at the first shot, with the ability to

expand and read more about it if they are interested in the song. Two chose two different presentations as equally favorable. Two people wanted the ability to have lesser information up front (figure 4/5) and then morph to give more information on request (figure 3/2). Six people chose the first layout, three chose the second, five chose the third, three chose the fourth, and four chose the fifth. Almost all said that they would definitely prefer to hear audio of the song. Twelve out of eighteen said that they would also like to see the video. People, felt much more strongly about video than they did about audio. They felt that video made the presentation much better than audio alone.
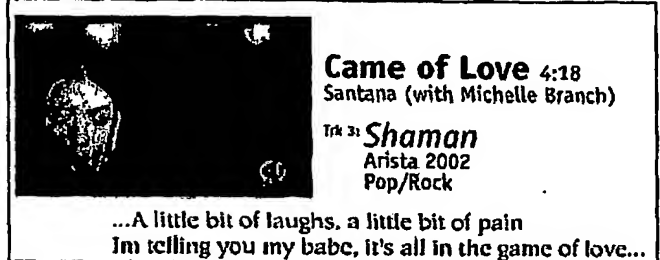

Came of Love 4:18
Santana (with Michelle Branch)

Trk 3: *Shaman*
Arista 2002
Pop/Rock

...A little bit of laughs, a little bit of pain
Im telling you my babe, it's all in the game of love...

Figure 1. Full summary horizontal.

...
A little bit of laughs
a little bit of pain
Im telling you my babe
It's all in the game of love
...

The Game of Love
Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18) Arista 2002: Pop/Rock

Figure 2. Full split summary.

The Game of Love
Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18) Arista 2002: Pop/Rock

...
A little bit of laughs
a little bit of pain
Im telling you my babe
It's all in the game of love
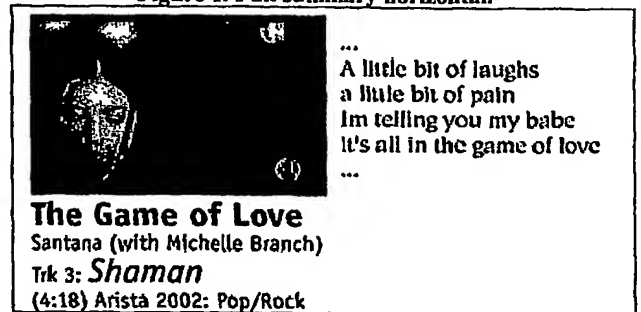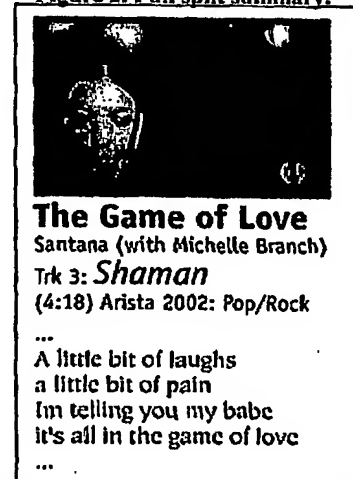...

Figure 3. Full summary in a vertical arrangement.

The Game of Love
Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18)

Figure 4. Short summary in a vertical arrangement.

The Game of Love
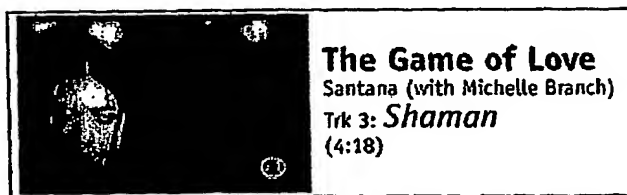Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18)

Figure 5. Short summary in a horizontal arrangement.

3. SYSTEM OVERVIEW

Here we describe the overall system architecture of the video music summarization. We assume that the system is receiving a video feed either from a broadcast/cable/satellite source, Internet streaming, or from a file stored in a video library. Also, we assume that connection to the Web is available in order to access song information such as title, artist, genre, and lyrics.

The general architecture is given in Figure 6. In our case, the video is digitized into MPEG-2 for storing and further use. Then the video is demultiplexed and separate audio, visual and transcript files are extracted. The transcript is extracted from the closed captions with time stamps inserted for each line. For these modalities we perform feature extraction: videotext detection, visual cuts, face detection, audio segmentation and classification, and transcript (closed captions) preprocessing. At this point all the features comprise a time stamped stream of data without any indication of song boundaries. Next, we determine the initial song boundary using the visual, auditory and textual features (see section 4.1). Next, using the initial boundaries and the transcript information, we determine the chorus location and chorus key phrases. Based on the chorus information, we use information from a Web site in order to find the title, name of the artist/performer, genre, and lyrics. The song boundary is then confirmed using the information about the exact song lyrics. We take into account that the lyrics on the Web site and the lyrics in the transcript do not always match perfectly. Based on the lyrics, we align the boundaries of the song using the initial boundary information and the lyrics. Alternatively, if transcript information is not available, the title page can be analyzed using OCR on the extracted videotext in order to find the artist name, song title, year, label information. Then Web information can be used to verify the output from the OCR step. With this information we can find the lyrics of the song from a Web site and perform our chorus detection method using textual information. However, we do not have time stamp information. Methods exist in the

literature for chorus detection in the audio domain, which can be applied in order to align the textual and the audio chorus.

Having the boundary for each song, and the audiovisual features we determine the best representative frames, and the best video clip for the song summary. The best representative frames include close-ups from the artist, the title image with the song information, artist, label, album, and year. Song summaries are stored in a song summary library. The users can access the program summaries and songs and summaries using a Web-based music video retrieval application called "Music Video Miner".
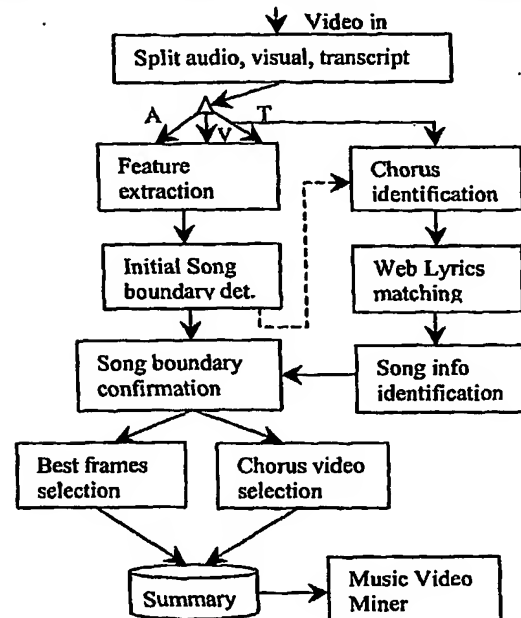

Figure 6. Overview of the music summarization system.

4. MUSIC VIDEOS SUMMARIZATION AND IDENTIFICATION

In order to summarize music videos, we looked at different sites on the web that sell CDs and offer samplers for viewers to hear before deciding to buy music. Almost all of them include the chorus of the song. Sometimes, they include the lead into the song. These audio samples are generally 30 seconds in duration on amazon.com and cduniverse.com. People remember the chorus of a song more than anything else as that is the part of the song that is heard most often. While guessing a title song, people usually do better if they hear the chorus rather than any other part of the song as that piece is heard most often in any song. It is made to be that the chorus is written such that it should not be too difficult. This was reinforced in the user needs analysis that the summary should definitely contain the chorus of the song.

Music video summarization is based on identification and summarization of individual songs. At a program level the summary consists of the list of songs. At the next level, each song consists of title, artist, and selected multimedia elements that represent the song.

4.1 Boundary Detection

There are two types of boundary problems present in music video summarization. The first one is to detect the song boundary automatically. The second problem is to detect the boundary of the chorus. As we explained in section 3, chorus and song boundary detection are intertwined and rely on each other.

We use audio, visual and transcript features. Visual features include: presence of videotext[9], face presence, abrupt cuts, color histograms[3].

Although faces are quite important for finding the main performing artist we have to note that music videos is one of the most challenging genres for video face detection. Very frequently the face presence is not detected because of special effects, lighting with various colors. Many faces are in a diagonal or horizontal position because people might be dancing, sleeping... Detection of videotext on the other hand is quite accurate because the intention of the producer is to make it easy to read and recognize. Presence of videotext at the beginning of the song helps delineate the boundaries between songs. Figure 7 shows face and text presence for 9000 frames of MTV video. The clip starts with a commercial break, then the song starts after five seconds, at frame 150, and lasts until frame 7300. Note that there is detected videotext from frame 76 to frame 91, and also from 361 to 406. This first text box is too small, and belongs to a commercial. The second series of text boxes contains the title information of the song. The text boxes are positioned at the low left portion of the screen. This title page of the song can be used as one indicator that the song has already started in order to determine the beginning of the song.

Cut changes are very frequent in music videos. In fact, our data shows that average cut distance is higher during a commercial break than during the songs. This is quite unusual since for most other genres, the commercial breaks exhibit lower cut distance than the program.

From the color change features we can infer the potential boundaries of the songs. Figure 8 shows the dominant color change in a 9 bin color quantization. The colors for the song "The Game of Love" are mostly in the dark gray range, and sometimes into yellowish range, because of the style of the video filming. The commercial break before the song until frame 150 and after the song, i.e. frame 7300 are using different colors. We are using the superhistogram method to infer the families of frames that exhibit similar colors. As reported earlier the same method can be used to infer the boundaries between programs. Music videos can be thought of as a small movie of their own, and this method is helpful in detecting the potential begin and end of songs.

In the audio domain we use audio segmentation and classification into multiple classes: 1) music, 2) speech, 3) speech with background music, 4) multiple people talking, 5) noise, 6) speech with noise, and 7) silence [14]. It is interesting to observe that in our feature analysis we see that melodic songs are correctly classified as belonging to the music category. However, for genres such as rap music the classification also shows speech during the song. Figure 9 shows the audio segmentation of the same video segment as for the previous two figures. The segment starts with speech and noise in the beginning with the real song starting at 150, where until 7300 the audio classification is showing mostly music. After frame 7300 the commercial break starts and we see segments belonging to different classes.

In order to determine the breaks we use the approximate boundaries from all the different features: videotext, superhistograms, audio, and transcript. Then we use the single descent approach through a stack of boundaries. Basically we use the fact that the transcript starts later than the visual and audio. From visual point of view we also get the videotext title page which normally appears after the start of the song. The begin boundary is then fine tuned with the superhistogram model for the song and the audio start for music classification. However if the title page is in a section classified as speech then the start time of either speech or speech with noise is sought out.
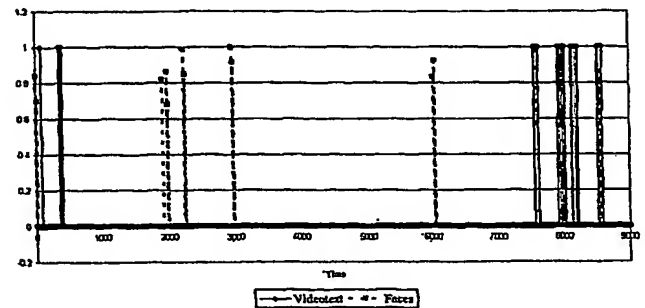

Figure 7. Face and text presence vs. time in frames.
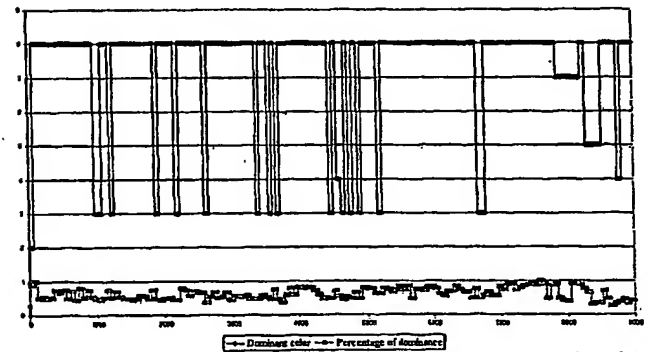

Figure 8. Dominant color values and amount of dominance.
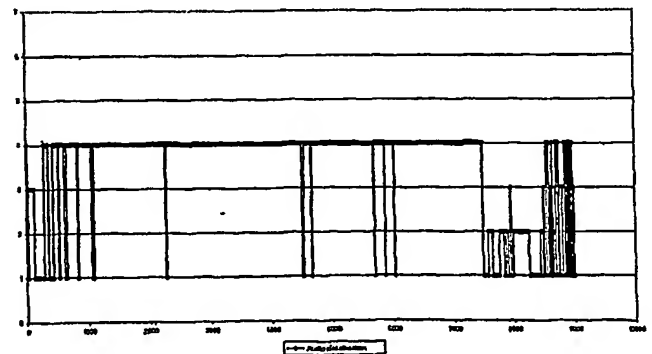

Figure 9. Audio segment classification.

4.2 Chorus Detection

In order to determine the chorus of a song, previous research has centered on music audio features. A common approach in order to find repeated segments in songs is to perform auto-correlation analysis. A chorus is repeated at least twice in popular songs. It is usually repeated thrice in most of the songs.

We decided to use the transcript (closed captions) in order to find the chorus of the song. The task is to detect the sections of the song that contains repeated words. Closed captions, however, are not perfect, and do contain a lot of typos, omissions etc. In order to recognize the chorus segments, the closed captions are processed in four steps consisting of, a) key-phrase detection, b) potential chorus detection, c) chorus candidate confirmation and d) irregular chorus detection and post analysis.

4.2.1 Keyphrase Identification

Chorus contains the lyrics in a song that are repeated most often. By detecting and clustering the phrases, we can identify the temporal location of the chorus segments. To select potential sections containing chorus we compile a tally (count) of phrases present in a song. These phrases are taken from the transcript and represent either a whole line of text on the television screen or parts of a line that have been broken up by delimiters such as a comma, period etc. As a new phrase is obtained, it is checked to see if the phrase exists in the tally. If it does, then the counter for that phrase is incremented. If not, a new bin is created for the new phrase and the counter is initialized to one. This process is repeated for all the text for each of the songs. At the end of the song, we adaptively select five to ten most frequently appearing phrases and designate these as keyphrases. The algorithms first starts with bins with two or more counts, and then keeps increasing the count threshold until we find less than 10 phrases which have a count more than the count threshold.

4.2.2 Candidate Chorus Detection

Potential candidates for a chorus segment are those that contain more than one occurrence of keyphrases. In order to find these segments, we find the timestamps at which each of the keyphrases occurs. For each timestamp of a keyphrase, a search is made to see if an existing potential chorus already has been detected. If the beginning of the potential chorus is within n seconds of the current timestamp, then the information about the chorus is modified to include this keyphrase. Based on an examination of a number of songs we work on the assumption here is that choruses are rarely more than 30 seconds long and $n=30$.

4.2.3 Chorus Candidate Confirmation

Only those candidates which contain three or more keyphrases are selected as choruses. A chorus is repeated at least twice in popular songs. It is usually repeated thrice in most of the songs. If more than three choruses are still left, then we select the three choruses that have the highest density of keyphrases. For example, if a chorus has eight keyphrases within 20 seconds as opposed to another having nine keyphrases in 17 seconds, then we choose the second over the first one.
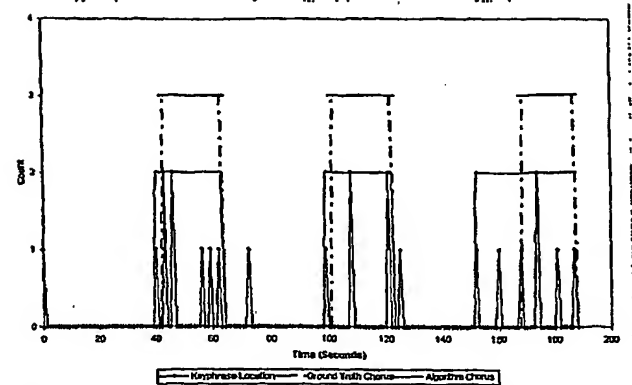

Figure 10 Keyphrase location, ground truth of chorus, and detected chorus

4.2.4 Irregular Chorus Detection and Post Analysis

For the summarization, we need to detect just one chorus correctly and identify the "key-chorus" among the choruses detected that will be presented to the users. There is a large variability within a song regarding the duration of different choruses. One chorus may be 15 seconds long and another one maybe 30 seconds long due to music etc. that is played during the chorus. This variability makes it hard to predict the location and length of choruses. We choose the chorus that is of medium length of the three choruses. We prefer the first chorus to the rest of the choruses as we hope to also get a "lead" into the song along with the first chorus. Also, the placement of chorus within a song is variable. So whereas the distance between beginning of first and second chorus is 50 seconds the distance between second and third chorus is 86 seconds in case of the song "Game of Love," by Santana. The final chorus analysis is used to select a chorus that has a reasonable distance from other choruses.

Figure 10 shows the location of the keyphrases in the song "Cry" by the artist Faith Hill depicted by a continuous line with dots. The ground truth of the choruses is depicted by the dotted line. The bold solid line presents the three choruses that were identified. We chose the first chorus to be included in the summary of the song because it satisfied all the above criteria.

4.2.5 Autocorrelation Analysis

In audio content analysis, researchers have used auto-correlation in order to find the chorus [10]. An autocorrelation analysis on the transcript can also be used to find the choruses. In order to find the autocorrelation function, we lay out all the words in the transcript in two dimensions and fill up the matrix with ones and zeroes depending on whether the words on both the dimensions are the same. Then we project this matrix diagonally and determine the peaks in this view, which now corresponds to choruses in the song. Figure 11 shows the autocorrelation matrix of the song. Figure 12 shows the result of autocorrelation analysis on the lyrics of the song "Game of Love." The song has 338 words. The peaks show the location of the choruses in the song. We can see in the autocorrelation matrix that there are three choruses.
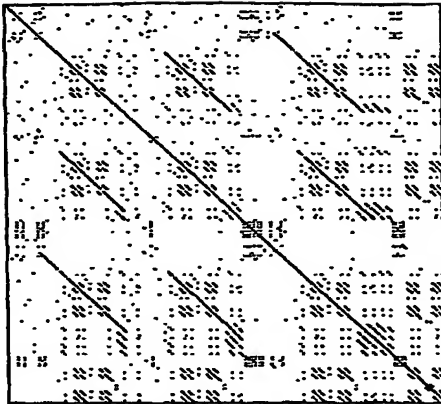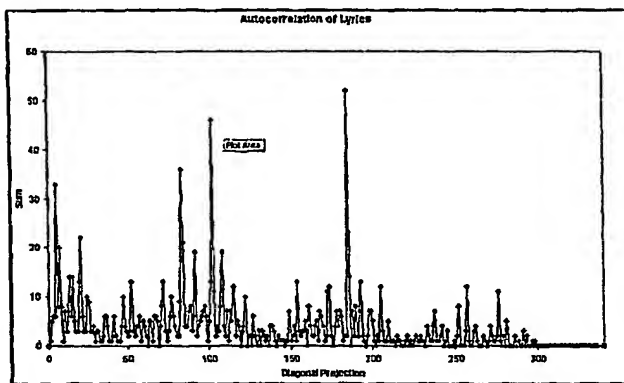
Figure 11 Lyrics Autocorrelation Matrix

Figure 12 Auto-correlation analysis result.

extracted for a set of music videos. Users were shown summaries of 10 songs as shown Figure 13. Clicking on the images in the summaries played the audio and video summary of the song, The users were asked to interact with summaries and then fill out our survey that had 27 questions in all. Four questions had subparts too, bringing the total number of actual questions to 38.

Tables 1 through 7 show the results of the survey that the users filled out after interacting with the summary. Table 1 Value of a music video summary. Table 2 Important elements of a music video summary. Table 3 Rank of media elements in order of importance in a summary. Table 4 Rank of text content elements order of importance in a summaryrespectively. And finally,

Table 7 Context: Where do users want the summary?, shows the context regarding where people want to view the survey. ·

We performed principal component analysis on the survey answers in order to uncover important trends. The analysis was broken into four parts as follows.

Figure 13 Interactive Music Video Summaries Interface

5. EXPERIMENTAL RESULTS

In order to evaluate our system we had to benchmark the automatic analysis as well as the user experience. For the automatic analysis, we present the results on the accuracy of song summarization algorithm in section 5.1. In section 5.2 we present the methodology we used to evaluate the music video summarization and also the results of user testing.

5.1 Summary Extraction Evaluation

We analyzed 4 hours of music videos that contain 38 songs. We analyzed the video and the closed captions to extract the summary of the songs. The most important parts of the summary are the identification of the title page from the music video and the chorus identification. The precision and recall for the identification of a title page from the song is 100%. We are able to determine correctly at least one frame that contains all the information about the song for all the 38 songs.

5.2 User Evaluation

Here we present our experimental results using VH1 video. We asked the eighteen people to perform user tests on the summaries

5.2.1 Part 1: Value

There is almost no variation in the responses here. There is a weak (eigenvalue = 1.3) connection among question 2,4,5,6, meaning that people giving a higher value in answer to one of that group tended to give a higher value to the others, too. This is surprising, as the sign is not reversed for 2: if you find summaries enjoyable, you also do not find they help you share! But the answers really don't show much of an effect.

5.2.2 Part 2: Elements

This set of questions is also very flat, with everyone tending to agree with everything except perhaps question 9 and 11. The analysis shows only two weak connection sets. One set, with eigenvalue = 2.1 is 9, 10, and -11 (that is, the question acts in the direction of "duration has value". That is people tended to see lyrics, chorus, and duration as being is a similar category, and voted the three of them in or out together. The second weak category, eigenvalue 1.6, is 9, 11, 13 (with duration having no value): people found lyric and video clip on one extreme and duration on the other. But given there are 10 questions here, not much is going on in this group either.

5.2.3 Part 3: Importance

Here we see some meaningful variation. People give uniformly high preference for audio, title, artist, chorus. Then, they strongly group together (eigenvalue = 5.8) text, lyrics, genre, beginning, music, and title page. This is a sort of "scholarly" factor, it seems: people either want them or don't. After that, there is a second weaker group (eigenvalue = 3.6) which makes the following choice: either close up, or video plus video segment. The last appreciable factor (eigenvalue = 2.1) trades off some interest in year and genre versus music and closeup; this probably getting into the noise. But you do have two good things to work with here: some people do want more scholarly detail and some don't. And some want full video while others go for a still.

5.2.4 Part 4: Where

About the only thing going on here is that everyone wants it on their PC, but only some want it elsewhere and when they do, they want it everywhere. That is, the eigenvalue of 4.7 groups together questions 22-26. There is a very weak factor (eigenvalue 1.3) that says people tend to link TV and stereo together and see them both as the opposite of the PDA, but again this might be noise.

Summarizing the above, what we arrive at is:

1) The summary should have audio, title, artist, and chorus.

2) Some people want the "scholar package" of text, lyrics, genre, beginning, music, and title page.

3) Some people want the closeup, but others want the video.

4) The summary should definitely play on the PC.

5) Some technophiles want to play everywhere, with the possible exception of the old technologies of TV and stereo.

6. APPLICATIONS

Figure 13 shows an application we have developed for browsing music videos. Using this interface, people can interactively search for music videos in the database by the name of an artist or group, the title of a song, or based on genre. The result of a search is shown in Figure 15.

There are different usage scenarios for this application for casual users, for music producers, or artists. For example, when preparing a play list for a party, a user can search for songs by browsing the music video summaries to decide what to include in the list. Music Video Miner can help in creating services for Music Videos on Demand as well as making music purchases.

Another scenario is to use the Music Video Miner coupled with automatic audio/video recommenders. Automatic recommender systems can use the information in the summary for clustering the music videos and selecting songs to compile a playlist and recommending new music to the user. Usually recommenders use high level information such as genre, artist title. Other recommenders use low level audio features. A recommender that uses both high level information as well as the extracted audio visual features, and chorus information has more in-depth information about the content.

Furthermore, when exploring new artists and domains of music based on collaborative filtering, a user still needs to apply his/her own personal filters. If all your friends can send you playlists,

there should be an effective way to sort through them, view them and decide what is important.

We envision many other applications such as music visualization, copyright infringement detection, tracking of content distribution recording user behavior and others. In visualization, the extracted features during music video summarization can serve as a basis for visualizing music videos and authoring novel multimedia presentations of the videos. Copyright infringement on the web can be made efficient by comparing summaries because they carry essential information in an abridged form. Content distribution tracking on the Web can also be made efficient if the information about the music video items is stored and compared based on the summaries instead of the full video.

Figure 14. Music Video Miner.

Figure 15 Music Video Miner Result Screen.

7. CONCLUSIONS

We have shown how a careful consideration of user needs, together with an efficient exploitation of the semantics of a tightly structured domain, has led to the creation and validation of a customizable user interface for browsing an extensive database of video content. Both investigations were critical. The user surveys and feedback determined the common patterns of user preferences, allowing a useful engineering compromise between full customization and design simplicity. We provide users with a choice of five basic slide presentations for the videos; each presents title, artist, chorus text, and chorus audio, but vary in other content and style. Based on an extensive survey, we

document that there appears to be only a small number of independent music video summary preferences: some users want a great deal of information, others very little (and few in between); some need access to the full video, others only an single closeup still; some will play it only on their PC platform, but others nearly everywhere else. The two-step semantic extraction and summarization process, based on the unique properties of music video and song structure, permitted a straightforward but user-pleasing compression of the content by a factor of approximately 10. We anticipate that other limited video genres, such as short movie reviews, sports highlight features, movie trailers, and other miniature genres can benefit from a similar approach, and may have similar results. We plan to pursue these, and hope by their study to come to illuminate those universal user-browsing preferences that may be shared in common by them.

8. REFERENCES

[1] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, and J. Martino, "CONIVAS: CONtent-based Image and Video Access System," presented at ACM Multimedia, Boston, 1996.

[2] L. Agnihotri, K. Devara, T. McGee, and N. Dimitrova, "Summarization of Video Programs Based on Closed Captioning", SPIE Conf. on Storage and Retrieval in Media Databases, San Jose, CA, January 2001, pp. 599-607.

[3] L. Agnihotri and N. Dimitrova, Video Clustering using superhistograms in large video archives, Visual 2000, Lyon, France November 2000

[4] A. Aner and J. R. Kender, "Video Summaries through Mosaic-Based Shot and Scene Clustering", In Proceedings European Conference on Computer Vision, Denmark, May 2002.

[5] M. A. Bartsch, Gregory H. Wakefield, To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing, IEEE Workshop on Apps. of Signal Proc to Acoustics and Audio, WASPAA, New Paltz, Oct 21-24, 2001.

[6] J.-Y. Chen, C. Taskiran, A. Albiol, E. J. Delp, and C. A. Bouman, "Vibe: A Compressed Video Database Structured for Active Browsing and Search," Purdue University 1999.

[7] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," ACM Multimedia, 1997.

[8] N. Dimitrova, H-J Zhang, B. Shahraray, I. Sezan, T. Huang, A. Zakhor, Applications of Video Content Analysis and Retrieval, IEEE Multimedia, Vol. 9, No. 3, Jul-Sept. 2002, pp. 42-55.

[9] N. Dimitrova, L. Agnihotri, C. Dorai, R Bolle, MPEG-7 VideoText Description Scheme for Superimposed Text, International Signal Processing and Image Communications Journal, September, 2000

[10] J. Foote, "Visualizing Music and Audio using SelfSimilarity ". In Proc. ACM Multimedia '99, pp. 77-80, Orlando, Florida, November 1999.

[11] M. Goto, "A Chorus-Section Detecting Method for Musical Audio Signals", ICASSP, Hong Kong, April 6-10, 2003.

[12] A. Gupta and R. Jain, "Visual Information Retrieval," Communications of the ACM, vol. 40, pp. 71-79, 1997.

[13] A. Hauptmann and M. Smith, "Text, speech, and vision for video segmentation: The Informedia project," presented at AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision, 1995.

[14] D. Li, I. K. Sethi, N. Dimitrova, McGee. Classification of General Audio Data for Content-Based Retrieval. Pattern Recognition Letters 2000.

[15] B. Logan and S. Chu, "Music summarization using keyphrases," in International Conference on Acoustics, Speech and Signal Processing, 2000.

[16] Yu-Fei Ma; Lie Lu; Hong-Jiang Zhang; Mingjing Li, "An Attention Model for Video Summarization," ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002.

[17] W. Niblack, X. Zhu, J. L. Hafner, T. Breuel, D. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S. I. Nin, S. Sull, B. Dom, B.-L. Yeo, S. Srinivasan, D. Zivkovic, and M. Penner, "Updates to the QBIC System," SPIE- Storage and Retrieval for Image and Video Databases VI, vol. 3312, pp. 150-161, 1998.

[18] G. Peeters, A. La Burthe, X. Rodet, Toward Automatic Music Audio Summary Generation from Signal Analysis, ISMIR 3rd International Conference on Music Information Retrieval , Paris, Oct. 13-17, 2002.

[19] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Movies Automatically," Journal on Visual Communications and Image Representation, vol. 7, pp. 345-353, 1996.

[20] H. Sundaram; L. Xie; S-F Chang, A Utility Framework for the Automatic Generation of Audio-Visual Skims , ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002.

Table 1 Value of a music video summary

		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	Music videos summaries allow me to quickly check out a list of songs and find something I want to play.				5	13

#						
2	The ability to email music video summaries to my friends does NOT help me to share the music I like.	4	10	2	2	
3	Music video summaries do NOT help me find songs I like.	3	12	3		
4	Music video summaries help me find songs I know.		1	1	8	8
5	Music video summaries make it easier to discover new artists.		1	2	7	8
6	Browsing and accessing music videos via summaries is enjoyable.		1		10	7

Table 2 Important elements of a music video summary

#		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
7	Seeing the artist's name makes it easy to find new songs by artist I like.			1	10	7
8	Seeing the artist name				10	8
9	Seeing lyrics in the music	1	2	6	6	3
10	Most songs are uniquely identifiable by their chorus.		1	2	5	10
11	Seeing the song duration adds NO value to the music video summary.	2	3	5	4	4
12	The ability to play an audio clip in the music video summary helps me find songs I am interested in.			1	8	9
13	The ability to play a video clip in the music video summary helps me find songs I am interested in.		1	2	7	8
14	Summary should allow me to identify the songs within 20 seconds.			1	8	9
15	Summary should include the chorus of a song.				6	6
16	The title screen of the music videos is important to see in the summary			1	11	6

Table 3 Rank of media elements in order of importance in a summary

Media Elements	Importance (1-5) (Least Imp – Most Imp)				
Audio				2	16
Video	2	1	5	5	5
Text	2	7	4	3	2

Table 4 Rank of text content elements order of importance in a summary

Text Content Elements	Importance (1-5) (Least Imp – Most Imp)				
Title of the song			2	2	14
Artist				1	17
Lyrics	1	3	7	4	3
Year	6	6	5		1
Track	11	3	3	1	
Duration	9	4	4	1	
Genre	6	4	5	3	

Table 5 Rank of audio elements order of importance in a summary

Audio Content Elements	Importance (1-5) (Least Imp – Most Imp)				
Chorus of the song				3	15
Beginning of the song	5	1	6	6	
Music from the song (without lyrics)	3	3	5	4	3

Table 6 Rank of video content elements order of importance in a summary

Video Content Elements	Importance (1-5) (Least Imp – Most Imp)				
Title page from video	2	1	7	3	5
Close up of artist	5	1	5	5	2
Video segment from the song	1	1	3	6	7

Table 7 Context: Where do users want the summary?

		Strongly Disagree	Dis-agree	Neu-tral	Agree	Strongly Agree
1	I do NOT want to access music video summaries on my PC.	8	10			
2	I want to access music video summaries on my portable MP3 player.	1	3	2	5	7
3	I want to access music video summaries on my Television.	1		1	6	10
4	I want to access music video summaries on my whole house stereo.	1		4	7	6
5	I want to access music video summaries on my PDA.	2	2	3	7	4
6	I want to access music video summaries on my Mobile phone.	2	3	5	5	3

Music Videos Miner

Lalitha Agnihotri
Dept. of Computer Science
Columbia University
New York, NY 10027
+1 212 939 7107

lalitha@cs.columbia.edu

Nevenka Dimitrova
Philips Research
345 Scarborough Road
Briarcliff Manor
+1 914 945 6059

nevenka.dimitrova@philips.com

John Kender
Dept. of Computer Science
Columbia University
New York, NY 10027
+1 212 939 7115

jrk@cs.columbia.edu

ABSTRACT

Music videos currently can be watched 24 hours a day via broadcast video on VH1, MTV1 and MTV2 and a variety of local channels. However, it is difficult to track all these channels and their offerings. People find it difficult if not impossible to find specific videos of interest. It is important to provide tools for browsing, searching and accessing music videos quickly. We present in this paper the digest of user-needs analysis we performed to find out what is important in music summaries. We demonstrate and evaluate a system that summarizes music videos and provides an interactive interface for browsing them. Starting with full music video programs, we segment out the individual song videos by finding their boundaries, which are distinguished by changes in color palette, in closed captions, and in frequency of shot transitions. The video summaries consist of list of song summaries. Each song summary consists of automatically selected high level information such as title, artist, duration, text of the chorus, as well as important audio and visual segments from the input video including the chorus as the most easily recognizable part of the song. Chorus locations are found noting patterns (autocorrelations) of repeated words and phrases in the lyrics. We propose a summarization framework based on content selection in different media based on a Bayesian Belief Network. We present the results from a user survey to evaluate the 1) value of the summary 2) content of the summary 3) context of the summary where and how the summary is viewed. We report that this summarization method yields high measurable user satisfaction. Based on the evaluation we designed and implemented a Web based application, called Music Video Miner, which allows people to retrieve music videos by artist, song, and genre.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]. Image/video retrieval

General Terms

Algorithms, Human Factors.

Keywords

Music video summarization, multimedia content analysis, user needs analysis, chorus detection, music databases.

1. INTRODUCTION

Music videos currently can be accessed via broadcast video on VH1, MTV1 and MTV2 24 hours a day. However, a viewer must abide by the broadcaster's terms in order to watch desired songs. A video recorder lets you record whole music shows, however there is no way to specify: I want to watch the song "You are the one" by Shania Twain even on advanced digital recorder products. Content analysis methods have been introduced in the literature that aim at providing high level access to specific parts of the program (e.g. highlights) [Scout]. Video summarization methods have been developed for news, sports, movies, sitcoms. While music content analysis is an active area of research, music videos analysis and summarization has been neglected amongst the existing work. We are presenting a system for browsing and searching music videos based on song summaries. Access to the individual song summaries is via a Web-based interface. Potential applications are numerous for music lovers, special interest groups, music producers, up and coming artists, as well as for copyright infringement detectors.

Video summarization has been a very active area of research [4][8]. However, music videos as a genre has not been investigated before. Reported content based retrieval systems include Query-By-Image-and Video-Content (QBIC) [17], VisualGrep [12], DVL of AT&T, InforMedia [13], VideoQ [7], MoCA [19], Vibe [6], and CONIVAS [1]. The InforMedia project is a digital video library system containing methods to create a short synopsis of each video primarily based on speech recognition, natural language understanding, and caption text. The MoCA project is designed to provide content-based access to a movie database. Besides segmenting movies into salient shots and generating an abstract of the movie, the system detects and recognizes title credits and performs audio analysis. Sundaram et al. [20]proposed a maximization utilization framework for creating audio-visual skims. Agnihotri et al. [2]introduced surface summarization of TV programs using transcript information. Aner et al. introduced mosaic-based scene representation that allows fast clustering of scenes into physical settings, as well as further comparison of physical settings across videos [4]. This enables

detection of plots of different episodes in situation comedies and serves as a basis for indexing whole video sequences. Ma et al. proposed an attention model that includes visuals, audio, and text modalities for summarization of videos [16].

On the other hand, music analysis and retrieval has only focused on the audio aspect [5]. Logan and Chu developed algorithms for finding key phrases in selections of popular music for audio thumbnailing [15]. Their method focused on the use of Hidden Markov Models and clustering techniques on mel-frequency cepstral coefficients (MFCCs), a set of spectral features that have been used with great success for applications in speech processing. Foote introduced audio "gisting," as an application of his measure of audio novelty [10]. This audio novelty score is based on the similarity matrix, which compares frames of audio based on features extracted from the audio. Foote leaves details such as the similarity metric and feature class as design decisions. Many of the previous methods detected as a chorus a repeated section of a given length and had difficulty identifying both ends of a chorus section and dealing with modulations i.e. music key changes. Recently the RefraiD method attempted to detect the chorus sections and estimate both ends of each section [11]. Peeters et al. derive dynamic features representing the time evolution of the energy content in various frequency bands [18]. Their approach is to consider the audio signal as a succession of "states" at various scales corresponding to the structure at various scales of a piece of music using unsupervised learning methods.

The paper is organized as follows. In section 2 we present the user needs analysis for finding a) what is the utility of music summaries and b) what is the important information in the summaries. In section 3 we present a system for automatic music video summarization. In section 4 we present the method of music segmentation and summarization. In section 5 we describe the experimental results and the user tests performed for evaluating the music video summarization method. Music Video Miner, a Web based application for music video browsing is presented in section 6. We conclude the paper in section 7.

2. USER NEEDS ANALYSIS

In order to ascertain the situational utility of music videos summaries, we decided to perform a user-needs analysis. Our test group consisted of eighteen people with ages range from 16 to 53 years. There were eleven women and seven men in our test group. The sessions were conducted one on one and were split into two parts. In the first part, a series of questions were asked which were modified based on the interest the person showed in a subject. After getting the users input on summarization, they were shown a series of summaries and asked to pick the one they liked most. People, could also offer alternate preferable representations if they did not like any of the summaries.

2.1 Q&A Session

The Q&A session was introduced by asking the group if they listen to music and if they watch music video channels (MVC). A series of questions were then asked. The participants were very forthcoming about their ideas and use of such a system. Everybody said that they could use such a system in their life that records and summarizes music videos. In the following section we present the different questions and a summary of the provided answers.

2.1.1 What do you like/do not like about MVCs?
Watching MVC's was almost universal. Viewership was, however, moderated by excessive talk, non-music related shows, or ill-matched choices on the MVC. Everybody enjoyed listening to music and watching the accompanying videos.

2.1.2 Would you like a tool that would allow access to a music videos library instead of watching MVC?
The participants were very excited about this idea and felt that it would be great to be able to watch music videos when they could.

2.1.3 Would you like to see a summary of the videos before viewing it?
All but one answered positively and enthusiastically to this question. Everybody felt that a little bit about the song would be good to see before deciding to watch the entire song. However, one participant felt that as long as the system retrieves all videos by a particular artist that has been requested, then he could just play it.

2.1.4 What should this music video summary contain?
Most answers included the title of the song along with the artist and the album the song came from. The year of the song should also be present in the summary. Seventeen out of eighteen preferred to hear the chorus of the song for the audio summary. Only one person said that the song piece does not matter. One participant said that beginning of the song also gives some idea about the song and would be nice to hear. The summary should include parts with singing rather than just music. The summary could depend on the genre. A shot of the performing artist or the band in the summary would be nice addition. The group felt that the video shown has to be somehow unique and particular to that song. However, for certain genre of songs, the group felt that there could be no distinctive part. Maybe the best person to select the video segment would be the director or marketing representative.

For most of the participants the length of the song is not an issue and hence does not need to be in the summary. Two people said that lyrics of the chorus are not important and should not be included in the summary.

2.1.5 Would you like to get additional information about the song, artist, etc?
There was a wide variation of opinions on this question. Some felt very strongly that they would never want any additional information about the song. Others felt that additional information should be presented as a hyperlink that the users could explore in case they decided that they like the song. Still other felt that additional information would be nice and should be included.

Links could be given for people to find songs in the same album or other songs by the artist/group. A link to discography of the artist was also mentioned. Information about where to buy the song was requested. Picture of CD cover alongside so that they can buy it when they go to a music store was mentioned several times. Statistics regarding how many copies of CD sold, the standing in Billboard charts, awards won would be interesting to see too. The director, location, the actors in the video were mentioned by a few participants who are into music videos.

2.1.6 How would you search such a music videos library?

Search by name of the artist and genre of the song came up most. Searching or retrieval by the most current songs was also a popular choice. Participants want to keep up to date with music and wanted the ability to sort by the most popular songs. They also want the ability to search based on themes, such as rainy videos, relaxing music etc. Ability to search by lyrics of a song they have heard or to search the songs that do not contain the title in the song is important. One person wanted to search for the video that shows a deer driving a car!

2.1.7 Where do you see a good use of such a system?

One recurrent theme for use of such a system was to create party lists. For people with big screen televisions, it is good to get a quick preview of audio and video in order to set up play lists for different parties (Christmas party play list, New Year party play list etc). Another theme was getting the top N songs. People said that they would like to get top 10 popular music videos at the end of the day. One person felt that it would be a good way of discovering new songs and for exploring new genres. People would like to use a system or finding the video that somebody recommended to them. They felt people could use the music videos retrieved by "dance" theme to learn new moves. Another market could be the karaoke kind of application. Here the lyrics have to be matched up to the closed captions.

Ability to get a music channel much like the satellite radio channels that is currently available so that only music videos are shown and there are no commercials and hosts was desired. Another scenario would be for a family where children could select their music while the parents search for their own music preference.

2.1.8 Other comments?

Participants felt that summaries should be different depending on whether they knew the song or not. If they already knew the song, then the title and artist are enough for them to venture into the song and the piece of the audio in the summary does not matter much. However, if they do not know the song, then more information is needed. Participants did not want excessive text, as it requires too much effort. A few people said that music videos are not very important. It is the song itself that is of more value to them. People did not want to pay for such a system.

2.2 User Interface Selection

Once the users answered the questions, they were shown different types of "results" that they might get when they search for a song in our system. The screen shots that were shown to the users are presented in figure 1-5. Once the users picked one out of the five versions, they were shown two more versions of the style that they liked. In one, the image was linked to audio and in the second the image was linked to audio and the video of the chorus of the song. The users then had to choose between the still image vs. audio vs. audio+video version that they would prefer to view. All but one participant had heard the song or recognized it after listening to the audio. Out of the five different presentations, all of them were equally selected. Most people did not have a strong preference about one presentation or other. However, some people preferred getting the shorter version at the first shot, with the ability to

expand and read more about it if they are interested in the song. Two chose two different presentations as equally favorable. Two people wanted the ability to have lesser information up front (figure 4/5) and then morph to give more information on request (figure 3/2). Six people chose the first layout, three chose the second, five chose the third, three chose the fourth, and four chose the fifth. Almost all said that they would definitely prefer to hear audio of the song. Twelve out of eighteen said that they would also like to see the video. People, felt much more strongly about video than they did about audio. They felt that video made the presentation much better than audio alone.
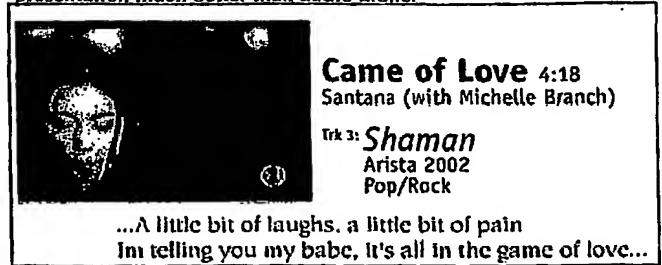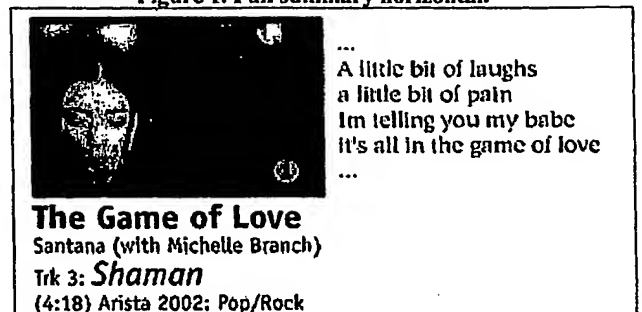

Came of Love 4:18
Santana (with Michelle Branch)

Trk 3: *Shaman*
Arista 2002
Pop/Rock

...A little bit of laughs, a little bit of pain
Im telling you my babe, It's all in the game of love...

Figure 1. Full summary horizontal.

...
A little bit of laughs
a little bit of pain
Im telling you my babe
It's all in the game of love
...

The Game of Love
Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18) Arista 2002: Pop/Rock

Figure 2. Full split summary.

The Game of Love
Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18) Arista 2002: Pop/Rock

...
A little bit of laughs
a little bit of pain
Im telling you my babe
it's all in the game of love
...

Figure 3. Full summary in a vertical arrangement.

Figure 4. Short summary in a vertical arrangement.

The Game of Love
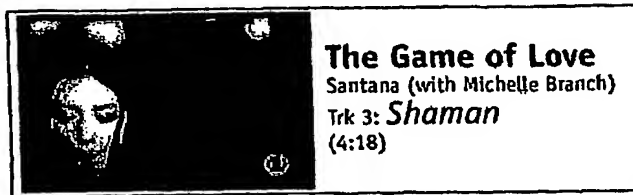Santana (with Michelle Branch)
Trk 3: *Shaman*
(4:18)

Figure 5. Short summary in a horizontal arrangement.

3. SYSTEM OVERVIEW

Here we describe the overall system architecture of the video music summarization. We assume that the system is receiving a video feed either from a broadcast/cable/satellite source, Internet streaming, or from a file stored in a video library. Also, we assume that connection to the Web is available in order to access song information such as title, artist, genre, and lyrics.

The general architecture is given in Figure 6. In our case, the video is digitized into MPEG-2 for storing and further use. Then the video is demultiplexed and separate audio, visual and transcript files are extracted. The transcript is extracted from the closed captions with time stamps inserted for each line. For these modalities we perform feature extraction: videotext detection, visual cuts, face detection, audio segmentation and classification, and transcript (closed captions) preprocessing. At this point all the features comprise a time stamped stream of data without any indication of song boundaries. Next, we determine the initial song boundary using the visual, auditory and textual features (see section 4.1). Next, using the initial boundaries and the transcript information, we determine the chorus location and chorus key phrases. Based on the chorus information, we use information from a Web site in order to find the title, name of the artist/performer, genre, and lyrics. The song boundary is then confirmed using the information about the exact song lyrics. We take into account that the lyrics on the Web site and the lyrics in the transcript do not always match perfectly. Based on the lyrics, we align the boundaries of the song using the initial boundary information and the lyrics. Alternatively, if transcript information is not available, the title page can be analyzed using OCR on the extracted videotext in order to find the artist name, song title, year, label information. Then Web information can be used to verify the output from the OCR step. With this information we can find the lyrics of the song from a Web site and perform our chorus detection method using textual information. However, we do not have time stamp information. Methods exist in the

literature for chorus detection in the audio domain, which can be applied in order to align the textual and the audio chorus.

Having the boundary for each song, and the audiovisual features we determine the best representative frames, and the best video clip for the song summary. The best representative frames include close-ups from the artist, the title image with the song information, artist, label, album, and year. Song summaries are stored in a song summary library. The users can access the program summaries and songs and summaries using a Web-based music video retrieval application called "Music Video Miner".
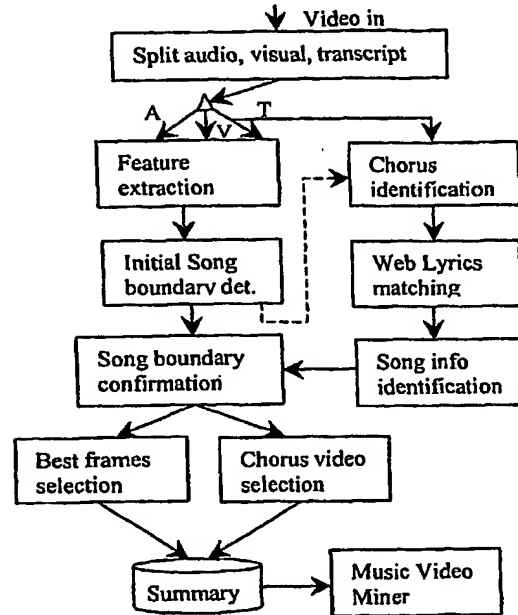

Figure 6. Overview of the music summarization system.

4. MUSIC VIDEOS SUMMARIZATION AND IDENTIFICATION

In order to summarize music videos, we looked at different sites on the web that sell CDs and offer samplers for viewers to hear before deciding to buy music. Almost all of them include the chorus of the song. Sometimes, they include the lead into the song. These audio samples are generally 30 seconds in duration on amazon.com and cduniverse.com. People remember the chorus of a song more than anything else as that is the part of the song that is heard most often. While guessing a title song, people usually do better if they hear the chorus rather than any other part of the song as that piece is heard most often in any song. It is made to be that the chorus is written such that it should not be too difficult. This was reinforced in the user needs analysis that the summary should definitely contain the chorus of the song.

Music video summarization is based on identification and summarization of individual songs. At a program level the summary consists of the list of songs. At the next level, each song consists of title, artist, and selected multimedia elements that represent the song.

4.1 Boundary Detection

There are two types of boundary problems present in music video summarization. The first one is to detect the song boundary automatically. The second problem is to detect the boundary of the chorus. As we explained in section 3, chorus and song boundary detection are intertwined and rely on each other.

We use audio, visual and transcript features. Visual features include: presence of videotext[9], face presence, abrupt cuts, color histograms[3].

Although faces are quite important for finding the main performing artist we have to note that music videos is one of the most challenging genres for video face detection. Very frequently the face presence is not detected because of special effects, lighting with various colors. Many faces are in a diagonal or horizontal position because people might be dancing, sleeping... Detection of videotext on the other hand is quite accurate because the intention of the producer is to make it easy to read and recognize. Presence of videotext at the beginning of the song helps delineate the boundaries between songs. Figure 7 shows face and text presence for 9000 frames of MTV video. The clip starts with a commercial break, then the song starts after five seconds, at frame 150, and lasts until frame 7300. Note that there is detected videotext from frame 76 to frame 91, and also from 361 to 406. This first text box is too small, and belongs to a commercial. The second series of text boxes contains the title information of the song. The text boxes are positioned at the low left portion of the screen. This title page of the song can be used as one indicator that the song has already started in order to determine the beginning of the song.

Cut changes are very frequent in music videos. In fact, our data shows that average cut distance is higher during a commercial break than during the songs. This is quite unusual since for most other genres, the commercial breaks exhibit lower cut distance than the program.

From the color change features we can infer the potential boundaries of the songs. Figure 8 shows the dominant color change in a 9 bin color quantization. The colors for the song "The Game of Love" are mostly in the dark gray range, and sometimes into yellowish range, because of the style of the video filming. The commercial break before the song until frame 150 and after the song, i.e. frame 7300 are using different colors. We are using the superhistogram method to infer the families of frames that exhibit similar colors. As reported earlier the same method can be used to infer the boundaries between programs. Music videos can be thought of as a small movie of their own, and this method is helpful in detecting the potential begin and end of songs.

In the audio domain we use audio segmentation and classification into multiple classes: 1) music, 2) speech, 3) speech with background music, 4) multiple people talking, 5) noise, 6) speech with noise, and 7) silence [14]. It is interesting to observe that in our feature analysis we see that melodic songs are correctly classified as belonging to the music category. However, for genres such as rap music the classification also shows speech during the song. Figure 9 shows the audio segmentation of the same video segment as for the previous two figures. The segment starts with speech and noise in the beginning with the real song starting at 150, where until 7300 the audio classification is showing mostly music. After frame 7300 the commercial break starts and we see segments belonging to different classes.

In order to determine the breaks we use the approximate boundaries from all the different features: videotext, superhistograms, audio, and transcript. Then we use the single descent approach through a stack of boundaries. Basically we use the fact that the transcript starts later than the visual and audio. From visual point of view we also get the videotext title page which normally appears after the start of the song. The begin boundary is then fine tuned with the superhistogram model for the song and the audio start for music classification. However if the title page is in a section classified as speech then the start time of either speech or speech with noise is sought out.
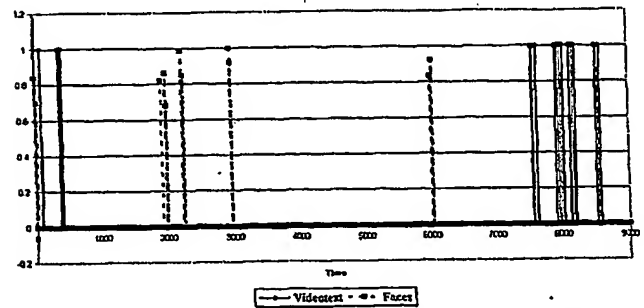

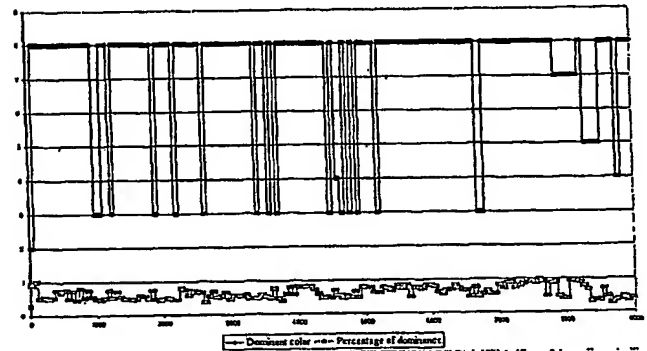**Figure 7. Face and text presence vs. time in frames.**

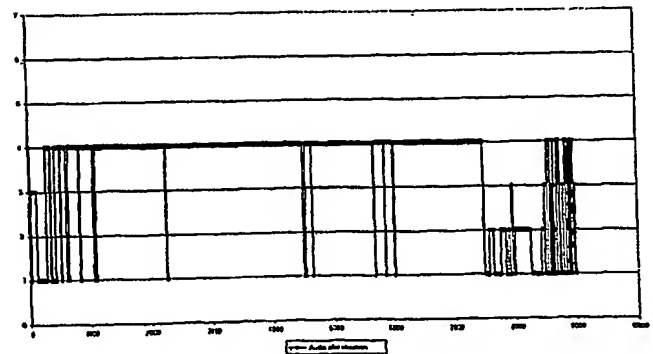**Figure 8. Dominant color values and amount of dominance.**

Figure 9. Audio segment classification.

4.2 Chorus Detection

'In order to determine the chorus of a song, previous research has centered on music audio features. A common approach in order to find repeated segments in songs is to perform auto-correlation analysis. A chorus is repeated at least twice in popular songs. It is usually repeated thrice in most of the songs.

We decided to use the transcript (closed captions) in order to find the chorus of the song. The task is to detect the sections of the song that contains repeated words. Closed captions, however, are not perfect, and do contain a lot of typos, omissions etc. In order to recognize the chorus segments, the closed captions are processed in four steps consisting of, a) key-phrase detection, b) potential chorus detection, c) chorus candidate confirmation and d) irregular chorus detection and post analysis.

4.2.1 Keyphrase Identification

Chorus contains the lyrics in a song that are repeated most often. By detecting and clustering the phrases, we can identify the temporal location of the chorus segments. To select potential sections containing chorus we compile a tally (count) of phrases. present in a song. These phrases are taken from the transcript and represent either a whole line of text on the television screen or parts of a line that have been broken up by delimiters such as a comma, period etc. As a new phrase is obtained, it is checked to see if the phrase exists in the tally. If it does, then the counter for that phrase is incremented. If not, a new bin is created for the new phrase and the counter is initialized to one. This process is repeated for all the text for each of the songs. At the end of the song, we adaptively select five to ten most frequently appearing phrases and designate these as keyphrases. The algorithms first starts with bins with two or more counts, and then keeps increasing the count threshold until we find less than 10 phrases which have a count more than the count threshold.

4.2.2 Candidate Chorus Detection

Potential candidates for a chorus segment are those that contain more than one occurrence of keyphrases. In order to find these segments, we find the timestamps at which each of the keyphrases occurs. For each timestamp of a keyphrase, a search is made to see if an existing potential chorus already has been detected. If the beginning of the potential chorus is within n seconds of the current timestamp, then the information about the chorus is modified to include this keyphrase. Based on an examination of a number of songs we work on the assumption here is that choruses are rarely more than 30 seconds long and $n=30$.

4.2.3 Chorus Candidate Confirmation

Only those candidates which contain three or more keyphrases are selected as choruses. A chorus is repeated at least twice in popular songs. It is usually repeated thrice in most of the songs. If more than three choruses are still left, then we select the three choruses that have the highest density of keyphrases. For example, if a chorus has eight keyphrases within 20 seconds as opposed to another having nine keyphrases in 17 seconds, then we choose the second over the first one.
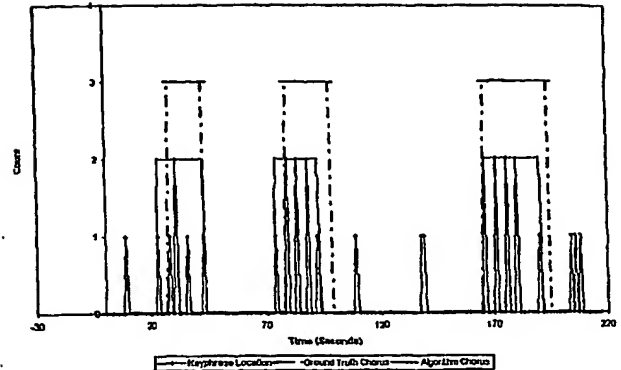

Figure 10 Keyphrase location, ground truth of chorus, and detected chorus

4.2.4 Irregular Chorus Detection and Post Analysis

For the summarization, we need to detect just one chorus correctly and identify the "key-chorus" among the choruses detected that will be presented to the users. There is a large variability within a song regarding the duration of different choruses. One chorus may be 15 seconds long and another one maybe 30 seconds long due to music etc. that is played during the chorus. This variability makes it hard to predict the location and length of choruses. We choose the chorus that is of medium length of the three choruses. We prefer the first chorus to the rest of the choruses as we hope to also get a "lead" into the song along with the first chorus. Also, the placement of chorus within a song is variable. So whereas the distance between beginning of first and second chorus is 50 seconds the distance between second and third chorus is 86 seconds in case of the song "Game of Love," by Santana. The final chorus analysis is used to select a chorus that has a reasonable distance from other choruses.

Figure 10 shows the location of the keyphrases in the song "Game of Love" depicted by a continuous line with dots. The ground truth of the choruses is depicted by the dotted line. The bold solid line presents the three choruses that were identified. We chose the first chorus to be included in the summary of the song because it satisfied all the above criteria.

4.2.5 Autocorrelation Analysis

In audio content analysis, researchers have used auto-correlation in order to find the chorus [10]. An autocorrelation analysis on the transcript can also be used to find the choruses. In order to find the autocorrelation function, we lay out all the words in the transcript in two dimensions and fill up the matrix with ones and zeroes depending on whether the words on both the dimensions are the same. Then we project this matrix diagonally and determine the peaks in this view, which now corresponds to choruses in the song. Figure 11 shows the autocorrelation matrix of the song. Figure 12 shows the result of autocorrelation analysis on the lyrics of the song "Game of Love." The song has 338 words. The peaks show the location of the choruses in the song. We can see in the autocorrelation matrix that there are three choruses.
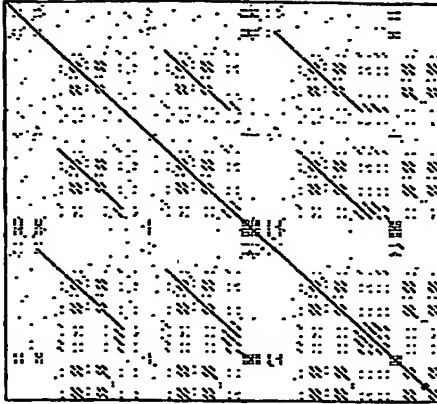
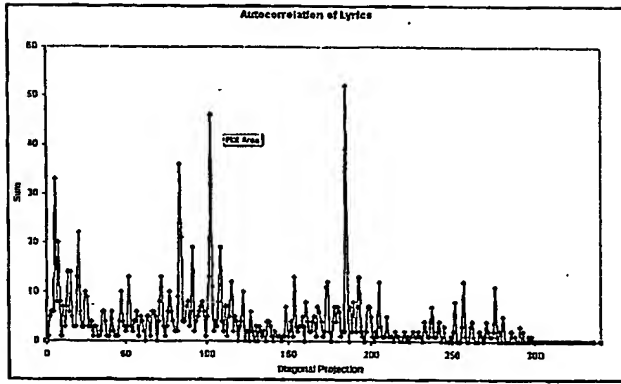**Figure 11 Lyrics Autocorrelation Matrix**

Figure 12 Auto-correlation analysis result.

4.3 Music Video Summary

A music video summary consists of content elements derived from the video in different media (audio, video, and text). We have considered using Bayesian Belief Networks to capture the generic content elements of a music video and Hidden Markov Models to capture the transitions of the music events and capture the structure of the composition. However, due to the large variability of the music creative process we think that computationally BBN is a more practical approach. For example, Abba's *Fernando* , has two parts: instrumental plus verse (V) and chorus (C). The order of musical events is V V C V C C. This is simple to model, however many songs have a bridging section between the chorus and the verse, and in many songs there is not even repeating chorus, but the whole song is one single monolithic verse. With the BBN approach even if one of the musical events is missing, we are still going to obtain a reasonable summary.

Figure 13 shows a BBN that can be used to model the function that is used to find the elements from the video that make up the summary.
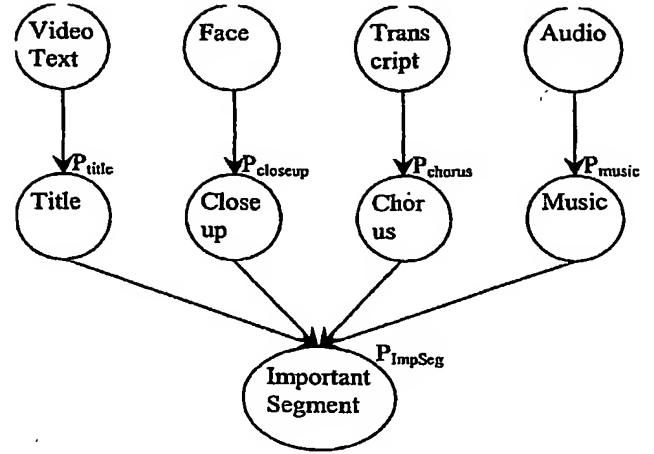

Figure 13 Bayesian Belief Network Model for Summarization

The probability for determining the important segment can be estimated as follows.

$$P(x \mid e^\wp) = \sum_{\wp_{mn}} P(x \mid \wp_{mn}) \prod \prod_{i=1}^{|\wp|} P(\wp_i \mid e)$$

$$P(x_i \mid e^\wp) = \sum_{\wp_{mn}} P(x_i, t, c, h, m)$$

$$= \sum_{\wp_{mn}} p(x_i \mid t)P(t)p(x_i \mid c)P(c)p(x_i \mid h)P(h)p(x_i \mid m)P(m)$$

Where $\wp = \{title, close-up, chorus, music\}$.

The value of m is 4 as we have 4 media elements. The value of n varies for each of the media elements depending on number of values that the probabilities can take. For example, for P(title) could be a value between 0 and 1 with steps of 0.1 depending the percent of screen covered with text. Thus n here is 10. Conceivably, we can include many more features, such as motion, audio-texture, lead instrument/singer highlight, in the parent nodes.

We have a selection criteria to decide the content to be presented in the summary for each of the media elements. The summary is the output from the selection functions that are defined as follows.

$$\psi_{Visual} P(x \mid e^\wp) = \begin{cases} 1 & if & \begin{aligned} P(x_i \mid e^\wp) \geq \theta_1 \\ \& P(vtext) \geq \theta_2 \\ \& P(face) \geq \theta_3 \end{aligned} \\ 0 & if & P(x_i \mid e^\wp) < \theta_1 \end{cases}$$

$$\psi_{Audio} P(x \mid e^\wp) = \begin{cases} 1 & if & \begin{aligned} P(x_i \mid e^\wp) \geq \theta_1 \\ \& P(music) \geq \theta_4 \end{aligned} \\ 0 & if & P(x_i \mid e^\wp) < \theta_1 \end{cases}$$

$$\psi_{Transcript} P(x \mid e^{b}) = \begin{cases} 1 & if & P(x_i \mid e^{b}) \geq \theta_1 \\ & & \& P(chorus) \geq \theta_s \\ 0 & if & P(x_i \mid e^{b}) < \theta_1 \end{cases}$$

The summary of a music video is a set consisting of the output of all the above selection functions:

$$S = \{\psi_{Audio}, \psi_{Video}, \psi_{Transcript}\}$$

In addition to addition to these elements derived from the video we add high level information, such as, artist, title, album extracted from the Web to complete the summary.

Of course, Bayesian Belief Network is just one way to model the selection of important elements for the summary. One can think of applying Sundaram's Utilization Maximization Framework or Ma's user attention model for summarization. These models are generative models for summarization. They model what the designer of the algorithm decides is important. Unsupervised machine learning techniques can be applied for music video visualization and summarization to find inherent structural patterns and highlights.

5. EXPERIMENTAL RESULTS

In order to evaluate our system we had to benchmark the automatic analysis as well as the user experience. For the automatic analysis, we present the results on the accuracy of song summarization algorithm in section 5.1. In section 5.2 we present the methodology we used to evaluate the music video summarization and also the results of user testing.

5.1 Summary Extraction Evaluation

We analyzed 4 hours of music videos that contain 38 songs. We analyzed the video and the closed captions to extract the summary of the songs. The most important parts of the summary are the identification of the title page from the music video and the chorus identification. The precision and recall for the identification of a title page from the song is 100%. We are able to determine correctly at least one frame that contains all the information about the song for all the 30 songs. For the chorus finder we analyzed the closed caption of the 27 unique songs contained in four hours of video. The recall and precision obtained for determining the choruses were 60% and 66% respectively.

5.2 User Evaluation

Here we present our experimental results using VH1 video. We asked the eighteen people to perform user tests on the summaries extracted for a set of music videos. Users were shown summaries of 10 songs as shown Figure 14. Clicking on the images in the summaries played the audio and video summary of the song, The users were asked to interact with summaries and then fill out our survey that had 27 questions in all. Four questions had subparts too, bringing the total number of actual questions to 38.

Tables 1 through 7 show the results of the survey that the users filled out after interacting with the summary. Table 1 gives the value of a music video summary. Table 2 shows the tally for the

important content elements of a music video summary. Table 3, gives the importance of different media elements. Table 4, Table 5, and Table 6 give the importance different content elements in text, audio and video media elements. And finally, Table 7, shows the context where the users want to view the summary.

We performed principal component analysis on the survey answers in order to uncover important trends. The analysis was broken into four parts as follows.
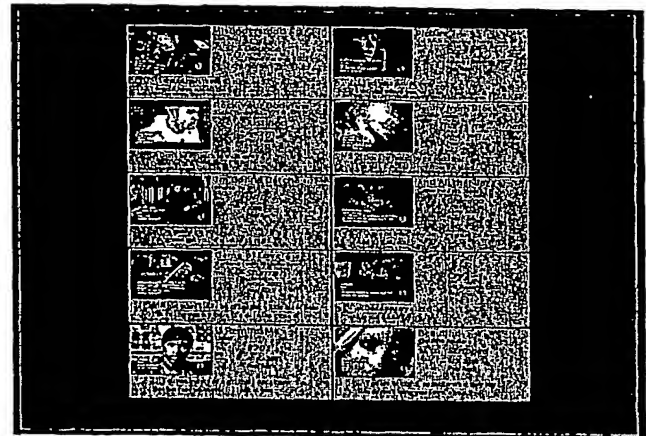

Figure 14 Interactive Music Video Summaries Interface

5.2.1 Part 1: Value

There is almost no variation in the responses here. There is a weak (eigenvalue = 1.3) connection among question 2,4,5,6, meaning that people giving a higher value in answer to one of that group tended to give a higher value to the others, too. One way to interpret the weak first factor is if people do like to use summaries to share, then they don't care much if summaries help them index into libraries, discover artists, or enjoy browsing. These people see summaries as a social exchange item rather than a tool.

5.2.2 Part 2: Elements

This set of questions is also very flat, with everyone tending to agree with everything except perhaps question 9 and 11. The analysis shows only two weak connection sets. One set, with eigenvalue = 2.1 is 9, 10, and -11 (that is, the question acts in the direction of "duration has value". That is people tended to see lyrics, chorus, and duration as being is a similar category, and voted the three of them in or out together. The second weak category, eigenvalue 1.6, is 9, 11, 13 (with duration having no value): people found lyric and video clip on one extreme and duration on the other. But given there are 10 questions here, not much is going on in this group either.

5.2.3 Part 3: Importance

Here we see some meaningful variation. People give uniformly high preference for audio, title, artist, chorus. Then, they strongly group together (eigenvalue = 5.8) text, lyrics, genre, beginning, music, and title page. This is a sort of "scholarly" factor, it seems: people either want them or don't. After that, there is a second weaker group (eigenvalue = 3.6) which makes the following

choice: either close up, or video plus video segment. The last appreciable factor (eigenvalue = 2.1) trades off some interest in year and genre versus music and closeup; this probably getting into the noise. But you do have two good things to work with here: some people do want more scholarly detail and some don't. And some want full video while others go for a still.

5.2.4 Part 4: Where

About the only thing going on here is that everyone wants it on their PC, but only some want it elsewhere and when they do, they want it everywhere. That is, the eigenvalue of 4.7 groups together questions 22-26. There is a very weak factor (eigenvalue 1.3) that says people tend to link TV and stereo together and see them both as the opposite of the PDA, but again this might be noise.

Summarizing the above, what we arrive at is:

1) The summary should have audio, title, artist, and chorus.

2) Some people want the "scholar package" of text, lyrics, genre, beginning, music, and title page.

3) Some people want the closeup, but others want the video.

4) The summary should definitely play on the PC.

5) Some technophiles want to play everywhere, with the possible exception of the old technologies of TV and stereo.

6. APPLICATIONS

Figure 13 shows an application we have developed for browsing music videos. Using this interface, people can interactively search for music videos in the database by the name of an artist or group, the title of a song, or based on genre. The result of a search is shown in Figure 16.

There are different usage scenarios for this application for casual users, for music producers, or artists. For example, when preparing a play list for a party, a user can search for songs by browsing the music video summaries to decide what to include in the list. Music Video Miner can help in creating services for Music Videos on Demand as well as making music purchases.

Another scenario is to use the Music Video Miner coupled with automatic audio/video recommenders. Automatic recommender systems can use the information in the summary for clustering the music videos and selecting songs to compile a playlist and recommending new music to the user. Usually recommenders use high level information such as genre, artist title. Other recommenders use low level audio features. A recommender that uses both high level information as well as the extracted audio visual features, and chorus information has more in-depth information about the content.

Furthermore, when exploring new artists and domains of music based on collaborative filtering, a user still needs to apply his/her own personal filters. If all your friends can send you playlists, there should be an effective way to sort through them, view them and decide what is important.

We envision many other applications such as music visualization, copyright infringement detection, tracking of content distribution recording user behavior and others. In visualization, the extracted features during music video summarization can serve as a basis for visualizing music videos and authoring novel multimedia presentations of the videos. Copyright infringement on the web

can be made efficient by comparing summaries because they carry essential information in an abridged form. Content distribution tracking on the Web can also be made efficient if the information about the music video items is stored and compared based on the summaries instead of the full video.
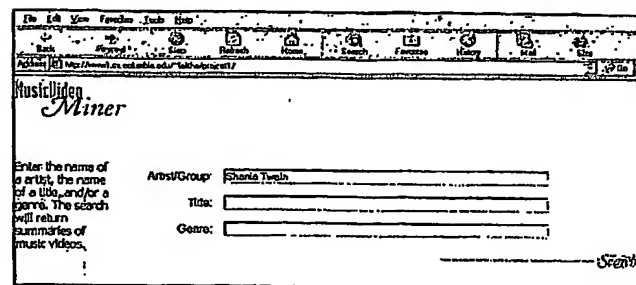

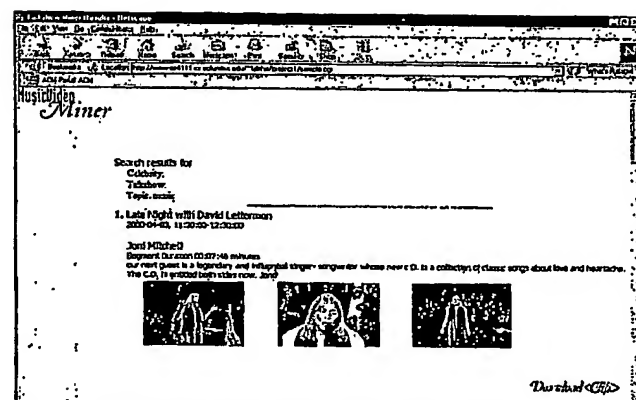**Figure 15. Music Video Miner.**

Figure 16 Music Video Miner Result Screen.

7. CONCLUSIONS

We have shown how a careful consideration of user needs, together with an efficient exploitation of the semantics of a tightly structured domain, has led to the creation and validation of a customizable user interface for browsing an extensive database of video content. Both investigations were critical. The user surveys and feedback determined the common patterns of user preferences, allowing a useful engineering compromise between full customization and design simplicity. We provide users with a choice of five basic slide presentations for the videos; each presents title, artist, chorus text, and chorus audio, but vary in other content and style. Based on an extensive survey, we document that there appears to be only a small number of independent music video summary preferences: some users want a great deal of information, others very little (and few in between); some need access to the full video, others only an single closeup still; some will play it only on their PC platform, but others nearly everywhere else. The two-step semantic extraction and summarization process, based on the unique properties of music

video and song structure, permitted a straightforward but user-pleasing compression of the content by a factor of approximately 10. We anticipate that other limited video genres, such as short movie reviews, sports highlight features, movie trailers, and other miniature genres can benefit from a similar approach, and may have similar results. We plan to pursue these, and hope by their study to come to illuminate those universal user-browsing preferences that may be shared in common by them.

8. REFERENCES

[1] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, and J. Martino, "CONIVAS: CONtent-based Image and Video Access System," presented at ACM Multimedia, Boston, 1996.

[2] L. Agnihotri, K. Devara, T. McGee, and N. Dimitrova, "Summarization of Video Programs Based on Closed Captioning", SPIE Conf. on Storage and Retrieval in Media Databases, San Jose, CA, January 2001, pp. 599-607.

[3] L. Agnihotri and N. Dimitrova, Video Clustering using superhistograms in large video archives, Visual 2000, Lyon, France November 2000

[4] A. Aner and J. R. Kender, "Video Summaries through Mosaic-Based Shot and Scene Clustering", In Proceedings European Conference on Computer Vision, Denmark, May 2002.

[5] M. A. Bartsch, Gregory H. Wakefield, To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing, IEEE Workshop on Apps. of Signal Proc to Acoustics and Audio, WASPAA, New Paltz, Oct 21-24, 2001.

[6] J.-Y. Chen, C. Taskiran, A. Albiol, E. J. Delp, and C. A. Bouman, "Vibe: A Compressed Video Database Structured for Active Browsing and Search," Purdue University 1999.

[7] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," ACM Multimedia, 1997.

[8] N. Dimitrova, H-J Zhang, B. Shahraray, I. Sezan, T. Huang, A. Zakhor, Applications of Video Content Analysis and Retrieval, IEEE Multimedia, Vol. 9, No. 3, Jul-Sept. 2002, pp. 42-55.

[9] N. Dimitrova, L. Agnihotri, C. Dorai, R Bolle, MPEG-7 VideoText Description Scheme for Superimposed Text,

[10] J. Foote, "Visualizing Music and Audio using SelfSimilarity ". In Proc. ACM Multimedia '99, pp. 77-80, Orlando, Florida, November 1999.

[11] M. Goto, "A Chorus-Section Detecting Method for Musical Audio Signals", ICASSP, Hong Kong, April 6-10, 2003.

[12] A. Gupta and R. Jain, "Visual Information Retrieval," Communications of the ACM, vol. 40, pp. 71-79, 1997.

[13] A. Hauptmann and M. Smith, "Text, speech, and vision for video segmentation: The Informedia project," presented at AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision, 1995.

[14] D. Li, I. K. Sethi, N. Dimitrova, McGee. Classification of General Audio Data for Content-Based Retrieval. Pattern Recognition Letters 2000.

[15] B. Logan and S. Chu, "Music summarization using keyphrases," in International Conference on Acoustics, Speech and Signal Processing, 2000.

[16] Yu-Fei Ma; Lie Lu; Hong-Jiang Zhang; Mingjing Li, "A User Attention Model for Video Summarization," ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002.

[17] W. Niblack, X. Zhu, J. L. Hafner, T. Breuel, D. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S. I. Nin, S. Sull, B. Dom, B.-L. Yeo, S. Srinivasan, D. Zivkovic, and M. Penner, "Updates to the QBIC System," SPIE- Storage and Retrieval for Image and Video Databases VI, vol. 3312, pp. 150-161, 1998.

[18] G. Peeters, A. La Burthe, X. Rodet, Toward Automatic Music Audio Summary Generation from Signal Analysis, ISMIR 3rd International Conference on Music Information Retrieval , Paris, Oct. 13-17, 2002.

[19] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Movies Automatically," Journal on Visual Communications and Image Representation, vol. 7, pp. 345-353, 1996.

[20] H. Sundaram; L. Xie; S-F Chang, A Utility Framework for the Automatic Generation of Audio-Visual Skims , ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002.

International Signal Processing and Image Communications Journal, September, 2000

Table 1 Value of a music video summary

		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	Music videos summaries allow me to quickly check out a list of songs and find something I want to play.				5	13
2	The ability to email music video summaries to my friends does NOT help me to share the music I like.	4	10	2	2	
3	Music video summaries do NOT help me find songs	3	12	3		

#		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	I like.					
4	Music video summaries help me find songs I know.		1	1	8	8
5	Music video summaries make it easier to discover new artists.		1	2	7	8
6	Browsing and accessing music videos via summaries is enjoyable.		1		10	7

Table 2 Important elements of a music video summary

#		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
7	Seeing the artist's name makes it easy to find new songs by artist I like.			1	10	7
8	Seeing the artist name				10	8
9	Seeing lyrics in the music	1	2	6	6	3
10	Most songs are uniquely identifiable by their chorus.		1	2	5	10
11	Seeing the song duration adds NO value to the music video summary.	2	3	5	4	4
12	The ability to play an audio clip in the music video summary helps me find songs I am interested in.			1	8	9
13	The ability to play a video clip in the music video summary helps me find songs I am interested in.		1	2	7	8
14	Summary should allow me to identify the songs within 20 seconds.			1	8	9
15	Summary should include the chorus of a song.				6	6
16	The title screen of the music videos is important to see in the summary			1	11	6

Table 3 Rank of media elements in order of importance in a summary

Media Elements	Importance (1-5) (Least Imp – Most Imp)				
Audio				2	16
Video	2	1	5	5	5
Text	2	7	4	3	2

Table 4 Rank of text content elements order of importance in a summary

Text Content Elements	Importance (1-5) (Least Imp – Most Imp)				
Title of the song			2	2	14
Artist				1	17
Lyrics	1	3	7	4	3
Year	6	6	5		1
Track	11	3	3	1	
Duration	9	4	4	1	
Genre	6	4	5	3	

Table 5 Rank of audio elements order of importance in a summary

Audio Content Elements	Importance (1-5) (Least Imp – Most Imp)

Chorus of the song				3	15
Beginning of the song	5	1	6	6	
Music from the song (without lyrics)	3	3	5	4	3

Table 6 Rank of video content elements order of importance in a summary

Video Content Elements	Importance (1-5) (Least Imp – Most Imp)				
Title page from video	2	1	7	3	5
Close up of artist	5	1	5	5	2
Video segment from the song	1	1	3	6	7

Table 7 Context: Where do users want the summary?

		Strongly Disagree	Dis-agree	Neu-tral	Agree	Strongly Agree
1	I do NOT want to access music video summaries on my PC.	8	10			
2	I want to access music video summaries on my portable MP3 player.	1	3	2	5	7
3	I want to access music video summaries on my Television.	1		1	6	10
4	I want to access music video summaries on my whole house stereo.	1		4	7	6
5	I want to access music video summaries on my PDA.	2	2	3	7	4
6	I want to access music video summaries on my Mobile phone.	2	3	5	5	3